

RESEARCH ARTICLE

VA-Index: Quantifying Assortativity Patterns in Networks with Multidimensional Nodal Attributes

Konstantinos Pelechrinis*, Dong Wei

School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, United States of America

* kpele@pitt.edu



OPEN ACCESS

Citation: Pelechrinis K, Wei D (2016) VA-Index: Quantifying Assortativity Patterns in Networks with Multidimensional Nodal Attributes. PLoS ONE 11(1): e0146188. doi:10.1371/journal.pone.0146188

Editor: Mansour Ebrahimi, Qom University, ISLAMIC REPUBLIC OF IRAN

Received: October 22, 2015

Accepted: December 14, 2015

Published: January 27, 2016

Copyright: © 2016 Pelechrinis, Wei. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data can be found at the following repository: <https://github.com/kpelechrinis/VAindex>.

Funding: This work was supported by Army Research Office (W911NF-15-1-0599) (KP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Network connections have been shown to be correlated with structural or external attributes of the network vertices in a variety of cases. Given the prevalence of this phenomenon network scientists have developed metrics to quantify its extent. In particular, the assortativity coefficient is used to capture the level of correlation between a single-dimensional attribute (categorical or scalar) of the network nodes and the observed connections, i.e., the edges. Nevertheless, in many cases a multi-dimensional, i.e., vector feature of the nodes is of interest. Similar attributes can describe complex behavioral patterns (e.g., mobility) of the network entities. To date little attention has been given to this setting and there has not been a general and formal treatment of this problem. In this study we develop a metric, the vector assortativity index (VA-index for short), based on network randomization and (empirical) statistical hypothesis testing that is able to quantify the assortativity patterns of a network with respect to a vector attribute. Our extensive experimental results on synthetic network data show that the VA-index outperforms a baseline extension of the assortativity coefficient, which has been used in the literature to cope with similar cases. Furthermore, the VA-index can be calibrated (in terms of parameters) fairly easy, while its benefits increase with the (co-)variance of the vector elements, where the baseline systematically over(under)estimate the true mixing patterns of the network.

Introduction

Assortativity mixing is a network phenomenon that describes the tendency of nodes to attach to others with similar characteristics. The mixing patterns are important in complex network theory since they can have many implications depending on the type of network examined. For instance, degree assortativity, that is, assortativity with respect to the node degree, is closely related with the resilience of a network to targeted attacks [1]. In the realm of social networks assortativity mixing with respect to external nodal attributes, usually termed as homophily [2], can reveal important information for the mechanisms that lead to friendship creation. As an illustrative example, studies of high school friendships have revealed a high degree of

homophily with respect to the students' race [3, 4], i.e., students tend to be friends with other students of the same race. The same tendency can be found in sexual relationship networks [5], while the marriage relationships exhibit assortativity mixing with respect to the age as well [6]. Spatial homophily, i.e., mixing with respect to locations visited by friends, has also been identified in social networks [7]. In the latter case, where a mutable attribute is examined, assortativity mixing can also be a sign of social influence, i.e., people first become friends—potentially due to irrelevant to the attribute examined reasons—and then they align their behavior with regards to the examined attribute. Of course, negative assortativity mixing, i.e., heterophily, can also be observed. For example, the sexual relationship social network is disassortative with regards to the gender of the nodes. The extent of this phenomenon has led to the integration of mixing patterns into generative network growth models [8, 9], while algorithms for recovering the underlying network connections exploit homophily as well [10, 11].

The central idea behind quantifying assortativity patterns in a network is to compare the number of edges that connect nodes of similar type with the expected number of these connections if the latter were picked at random. For example, if every node i is associated with a scalar value x_i (e.g., its age), we can compute the normalized covariance of the values x_i and x_j at the ends of an edge $\{i, j\}$ and then the assortativity coefficient r is given by [1, 6]:

$$r = \frac{\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j}{\sum_{ij} \left(k_i \delta_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j} \quad (1)$$

where \mathbf{A} is the adjacency matrix of the network, k_i is the degree of node i , m is the number of edges in the network and δ_{ij} is the Kronecker's delta. The values of r are bounded between $[-1, 1]$ —in practice the minimum value is $-1 \leq r_{min} < 0$ depending on the number of different node types [6]—and hence, allows for relative comparison between different networks and/or attributes. An alternative approach that can be used to quantify the levels of homophily is to include the attribute under examination (i.e., x_i) as a regressor in a model for network relations [12]. This will allow us to evaluate the statistical significance of specific variables in the formation of the network. This approach is different in the sense that while it can provide us with an estimation of the statistical importance of the corresponding attribute in the network formation it does not provide us with a fine-grained view. In particular, the regression coefficient is not bounded within a specific range and hence, direct inter-network and/or inter-attribute comparisons can be challenging.

While metrics for quantifying the assortativity mixing with respect to enumerative or scalar attributes have been developed, formal treatment of mixing patterns for vector nodal attributes has not received much of attention [13]. Nevertheless vector attributes appear in a variety of settings. In directed networks, the full degree information for a node is represented through a two-dimensional vector each element of which represents the in and out degree. Hence, if we do not want to lose the direction information, the degree assortativity needs to consider a vector rather than a scalar attribute [14]. Vector attributes can also describe behavioral aspect of nodes in social networks. For instance, the urban mobility of a city-dweller can be described through a vector each element of which captures the different types of locations he visits. Similarly, reviewers/buyers on electronic markets such as Amazon can be associated with a vector that captures their behavior with regards to the types of objects they are reviewing/buying. Furthermore, the analysis of composite networks that consist of multiple types of nodes and/or edges, requires novel metrics even for the simple scenario of the degree assortativity. In this setting, the degree of a node is not a single number anymore, but rather a vector based on the different types of edges attached to the node. Hence, formally put in this work we are interested in the following problem:

Problem 1 Given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ ($|\mathcal{V}| = n$ and $|\mathcal{E}| = m$), where node $v \in \mathcal{V}$ is associated with a vector $\mathbf{x}_v \in \mathbb{R}^q$ estimate the assortativity $r \in [-1, 1]$ of \mathcal{G} with regards to vectors \mathbf{x}_i , $i \in \mathcal{V}$.

As alluded to above, the requirement that $r \in [-1, 1]$ will allow us to directly compare the assortativity of different networks and/or different attributes. While recent studies have dealt with specific instances of Problem 1 the literature is still missing a formal metric that is generally applicable and can then be adopted to specific cases. For instance, Foster *et al.* [14] define 4 different types of degree assortativity in a directed network in order to account for the two different degree types (i.e., in and out). Block and Grund [15] examine the network dynamics of a friendship network when individuals have an increasing number of attributes in common. They utilize stochastic actor-oriented models and they show that there appears to be a diminishing effect with the number of common attributes. However, their approach is applicable only to longitudinal and directed network data. In a slightly different direction, Sánchez *et al.* [16] develop a method for the statistical selection of congruent subspaces, i.e., multivariate subspaces that have high dependency with the network structure. They further show that their method enhances outlier detection. Pelechrinis [17] recently developed a generic method that can provide an answer to Problem 1 in its generic form. Nevertheless, the proposed method is based on clustering the vector attributes of the network nodes. Given that clustering is known to be an ill-posed problem, at least under certain axiomatic frameworks [18, 19], selecting an appropriate clustering algorithm for all cases might be hard if not impossible and hence, the practical applicability of this work is limited. Despite the aforementioned efforts for tackling directly the multi-dimensional assortativity, the majority of the literature that deals with similar problems treats every element of the vector feature in isolation (e.g., [20, 21]). A similar approach will also for our baseline metric for comparison.

In this work we introduce a novel network metric, which we call *VA-index*, for quantifying the multi-dimensional assortativity. In a nutshell, our metric is based on network randomization and empirical hypothesis testing (see [Materials and Methods](#)). We evaluate our method by utilizing synthetic network datasets and comparing it with a baseline metric from existing literature (see [Results](#)). Finally, we discuss the significance and the implications of the proposed metric (see [Discussion](#)).

Materials and Methods

In order to solve Problem 1 we develop *VA-index*, whose computation combines network randomization with statistical hypothesis testing. In a nutshell, the intuition of our approach is based on comparing the pairwise average similarity of the vector attributes \mathbf{x} of connected nodes in \mathcal{G} with the one expected if connections were made at random. The distribution for the *randomly* expected average similarity can be estimated through Monte Carlo simulations of network randomizations. The latter can be either fully random (i.e., Erdős-Rényi random networks [22]) or control for specific network properties such as the degree distribution (e.g., configuration model [25]) and/or even for external properties (e.g., home location of users in a social network). We further perform a hypothesis test to evaluate the statistical significance of any difference observed, while we transform the observed effect size to a value bounded between $[-1, 1]$ through the standardized mean difference. In more detail *VA-index* computation comprises of the following steps:

Step 1. We first calculate the average pairwise similarity of connected nodes in \mathcal{G} with respect to the attribute vectors \mathbf{x} . Given a pair of nodes $v, u \in \mathcal{V}$ connected in \mathcal{G} , with attribute vectors \mathbf{x}_v and \mathbf{x}_u their similarity is $\xi(\mathbf{x}_v, \mathbf{x}_u)$, where ξ is a similarity measure in \mathbb{R}^q . Then the

average pairwise similarity of connected nodes in \mathcal{G} is:

$$\bar{\xi}_{\mathcal{G}} = \frac{\sum_{(v,u) \in \mathcal{E}} \xi(\mathbf{x}_v, \mathbf{x}_u)}{m} \quad (2)$$

Step 2. At this step we bootstrap through Monte Carlo simulations the estimation of the average pairwise similarity of connected nodes if these connections were made at random, $\bar{\xi}_{rand}$. In particular, we re-shuffle all m edges of \mathcal{G} uniformly—or controlling for other parameters—at random and generate \mathcal{B} randomized network structures (S1 Text). For each of the randomized networks i we calculate the average pairwise similarity of (randomly) connected nodes, $\bar{\xi}_i$. Hence, we get a sample $\Xi = \{\bar{\xi}_i | 1 \leq i \leq \mathcal{B}\}$, which essentially provides us with an estimate for the probability distribution of $\bar{\xi}_{rand}$, $f(\bar{\xi}_{rand})$.

Step 3. At this step we will examine where $\bar{\xi}_{\mathcal{G}}$ lays with respect to $f(\bar{\xi}_{rand})$ in order to identify whether there is positive, negative or random mixing in the network with respect to vector attributes \mathbf{x} . More specifically, we will examine the quantile of $f(\bar{\xi}_{rand})$ that includes the value of $\bar{\xi}_{\mathcal{G}}$. For example, in Fig 1 we present the probability distribution $f(\bar{\xi}_{rand})$ along with the 95% confidence interval

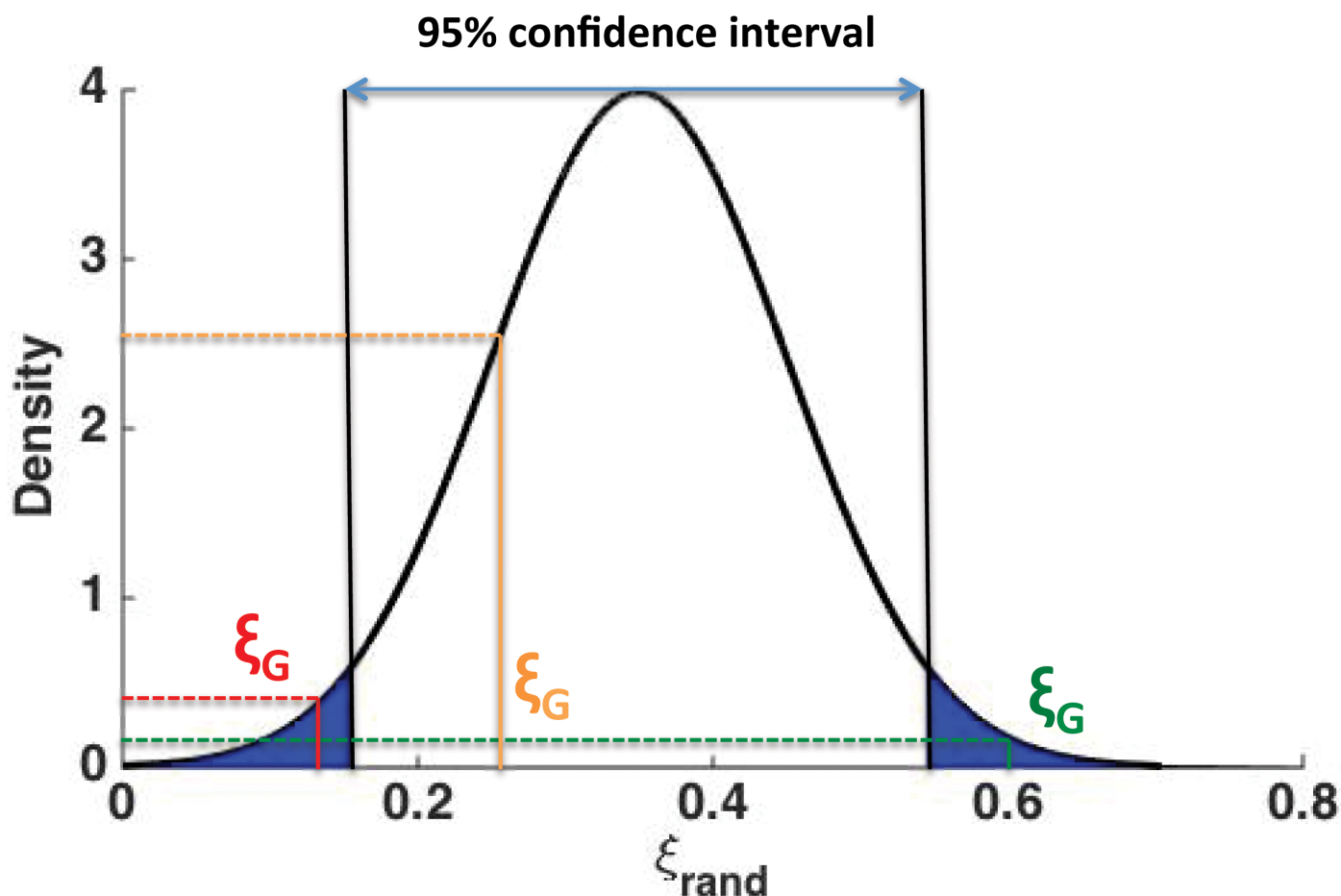


Fig 1. Decision boundaries for positive, negative or random mixing in the network. If the average similarity of connected nodes in the network $\bar{\xi}_{\mathcal{G}}$ falls in the top 2.5% quantile of $f(\bar{\xi}_{rand})$ (e.g., green line) we can conclude—at the significance level of $\alpha = 0.05$ —that the network is positively mixed. Similarly, if $\bar{\xi}_{\mathcal{G}}$ falls in the bottom 2.5% quantile of $f(\bar{\xi}_{rand})$ (e.g., red line) the network is negatively mixed. Otherwise (e.g., orange line) we cannot reject the hypothesis that the network is randomly mixed with respect to \mathbf{x} .

doi:10.1371/journal.pone.0146188.g001

confidence interval for $\bar{\xi}_{rand}$ (as computed from the \mathcal{B} randomized networks). There are now three possibilities for $\bar{\xi}_{\mathcal{G}}$:

- $\bar{\xi}_{\mathcal{G}}$ falls in the bottom 2.5% quantile of $f(\bar{\xi}_{rand})$. In this case the average similarity of connected nodes in \mathcal{G} is significantly smaller (at the significance level of $\alpha = 0.05$) than what we would have expected if connections were made at random. Hence, \mathcal{G} is negatively mixed with respect to \mathbf{x} .
- $\bar{\xi}_{\mathcal{G}}$ falls in the top 2.5% quantile of $f(\bar{\xi}_{rand})$. In this case the average similarity of connected nodes in \mathcal{G} is significantly larger (at the significance level of $\alpha = 0.05$) than what we would have expected if connections were made at random. Hence, \mathcal{G} is positively mixed with respect to \mathbf{x} .
- $\bar{\xi}_{\mathcal{G}}$ falls within the 95% confidence interval of $f(\bar{\xi}_{rand})$. In this case we cannot reject (at the significance level of $\alpha = 0.05$) the hypothesis that \mathcal{G} is randomly mixed with respect to \mathbf{x} .

Note that the above process is essentially the result of the following hypothesis test:

$$H_0 : \bar{\xi}_{rand} = \bar{\xi}_{\mathcal{G}} \quad (3)$$

$$H_1 : \bar{\xi}_{rand} \neq \bar{\xi}_{\mathcal{G}} \quad (4)$$

We would like to emphasize here that we do not perform a t-test (or any other standardized, off-the-shelf, hypothesis test), since we can directly estimate the empirical probability distribution $f(\bar{\xi}_{rand})$ from the Monte Carlo simulations and hence, obtain an empirical p-value (alternatively the corresponding confidence intervals).

Step 4. At this final step we quantify the levels of assortativity mixing in the network by comparing $\bar{\xi}_{\mathcal{G}}$, with the mean of the sample Ξ , m_{Ξ} . In particular, we first calculate the standardized mean difference as follows:

$$d = \frac{\bar{\xi}_{\mathcal{G}} - m_{\Xi}}{\sigma_{rand}} \quad (5)$$

where σ_{rand} is the expected standard deviation of the pairwise similarity in the randomized network, which can be calculated through the repeated randomizations. Then we transform this standardized difference to a value bounded between -1 and 1, which is our final VA-index α , through the following transformation:

$$\alpha = \frac{d}{\sqrt{d^2 + \epsilon}} \quad (6)$$

The final output is the VA-index α from step 4 as well as the empirical p-value for this index (at the significance level $\alpha = 0.05$) obtained through step 3. Note here that, by choosing different quantiles in step 3, we can perform the same test at a different significance level. Furthermore, the value of ϵ used in Eq (6) will be an evaluation parameter of the VA-index. Fig 2 summarizes the above steps.

In order to evaluate the above method we will rely on synthetic network data for which we know the ground truth with regards to the mixing patterns (S2 Text). However, prior to presenting our evaluations we would like to emphasize on the fact that while the VA-index is inspired by the assortativity coefficient, it is not a direct generalization of it in higher dimension. Both metrics' key idea is comparing features of the real network (i.e., number of edges between nodes with similar attributes in the case of assortativity coefficient and the average

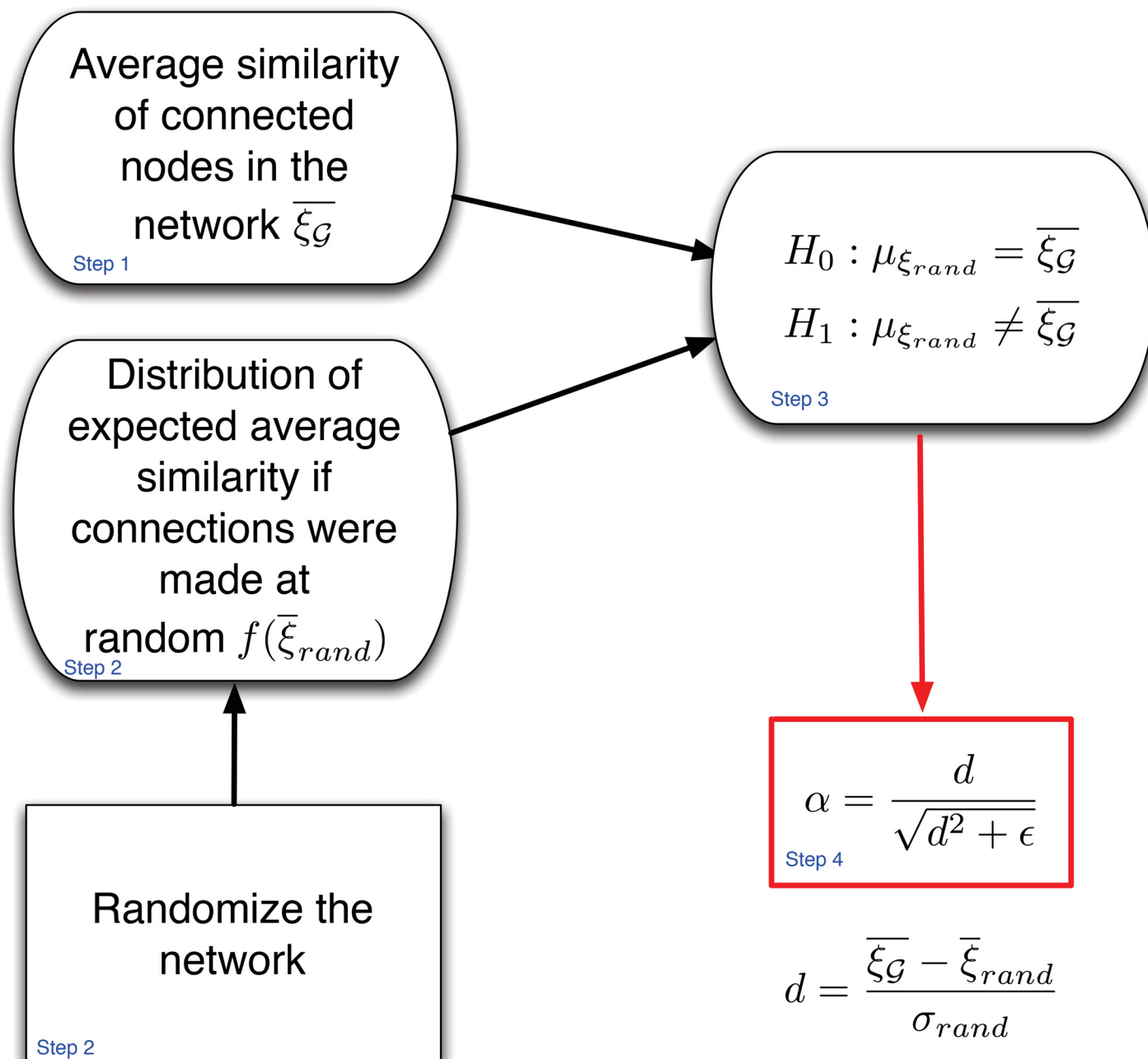


Fig 2. The computation of VA-index in a nutshell. VA-index involves network randomization and empirical hypothesis testing for quantifying the assortativity of a network with respect to a multi-dimensional nodal attribute.

doi:10.1371/journal.pone.0146188.g002

similarity of connected nodes for VA-index) with a randomized version of it. Note here that, the assortativity coefficient is based on comparisons with the Erdős-Rényi random graph model, while the VA-index can adopt other randomized generative models as well. For an appropriate choice of similarity metric and normalization (step 4), the VA-index can potentially recover the assortativity coefficient exactly. In particular, given that the assortativity coefficient is normalized using the maximum possible modularity (denominator of Eq (1)), we would need to normalize the VA-index with the maximum possible average similarity that

can be observed in the network analyzed. The latter is extremely hard—if not intractable—to obtain analytically in the general case, while it is computationally expensive to compute it through Monte Carlo simulations.

Results

We compare our system with a baseline extension of the assortativity coefficient. In particular, we calculate the assortativity coefficient r_i for each element i of \mathbf{x} . Our baseline assortativity is then given by:

$$r_{base} = \frac{\sum_{i=1}^q r_i}{q} \quad (7)$$

Given that in our synthetic data we know the actual assortativity patterns of the network our evaluation metric is the Root Mean Square Error (RMSE) of the assortativity values obtained from the VA-index and the baseline. More specifically, we will compare the RMSE of the VA-index and the baseline, while we will also examine the sensitivity of the VA-index with respect to parameters such as the similarity metric ξ used and the value of ϵ in Eq (5).

Sensitivity to ϵ and ξ

We begin by evaluating the performance of the VA-index with respect to the choice of ϵ and ξ . In particular, we consider three different similarity functions, namely cosine similarity, correlation similarity and a Euclidean distance-based similarity (S3 Text). Fig 3 depicts our results as a function of the value of ϵ in Eq (5).

As we can see the performance is very similar regardless of the specific similarity metric used. Furthermore, the RMSE error is much lower compared to the baseline for a wide range of values of ϵ . These results imply that the VA-index is not sensitive to the choice of ξ and ϵ , even though a suggested configuration appears to be the cosine similarity with a value of $\epsilon = 1$.

Comparison with the baseline

We now compare the VA-index α with the baseline assortativity coefficient r_{base} and evaluate the performance based on different levels of variance s , correlation c and density δ of Σ (S2 Text). The left part of Fig 4 depicts the results with regards to variance s , while all the VA-index results presented are obtained with cosine distance and $\epsilon = 1$. As we can see for low levels of variance, the two methods perform equally well. However, with an increase in the variance of the elements of the nodal vector attribute \mathbf{x} , the VA-index clearly outperforms the baseline with respect to the achieved RMSE. An increased variance at the vector elements leads the baseline coefficient to systematically make erroneous estimations for each dimension, which add-up at the end. In contrast, the VA-index considers all the elements of the vector simultaneously and hence, alleviates these problems. Similarly, our method outperforms the baseline metric regardless of the correlation between the elements of \mathbf{x} or the fraction of non-zero off-the-diagonal elements of Σ .

Next we compare the absolute errors made by the VA-index and the baseline. In particular, with $r_{true,v}$ being the true assortativity of network v , $r_{base,v}$ being the assortativity obtained from the baseline method and α_v being the VA-index, we calculate:

$$\Delta e_v = |r_{true,v} - \alpha_v| - |r_{true,v} - r_{base,v}| \quad (8)$$

A value of $\Delta e_v < 0$ implies that the VA-index can better recover the ground truth assortativity of a network. Hence, for every synthetic network we create we calculate Δe and perform a

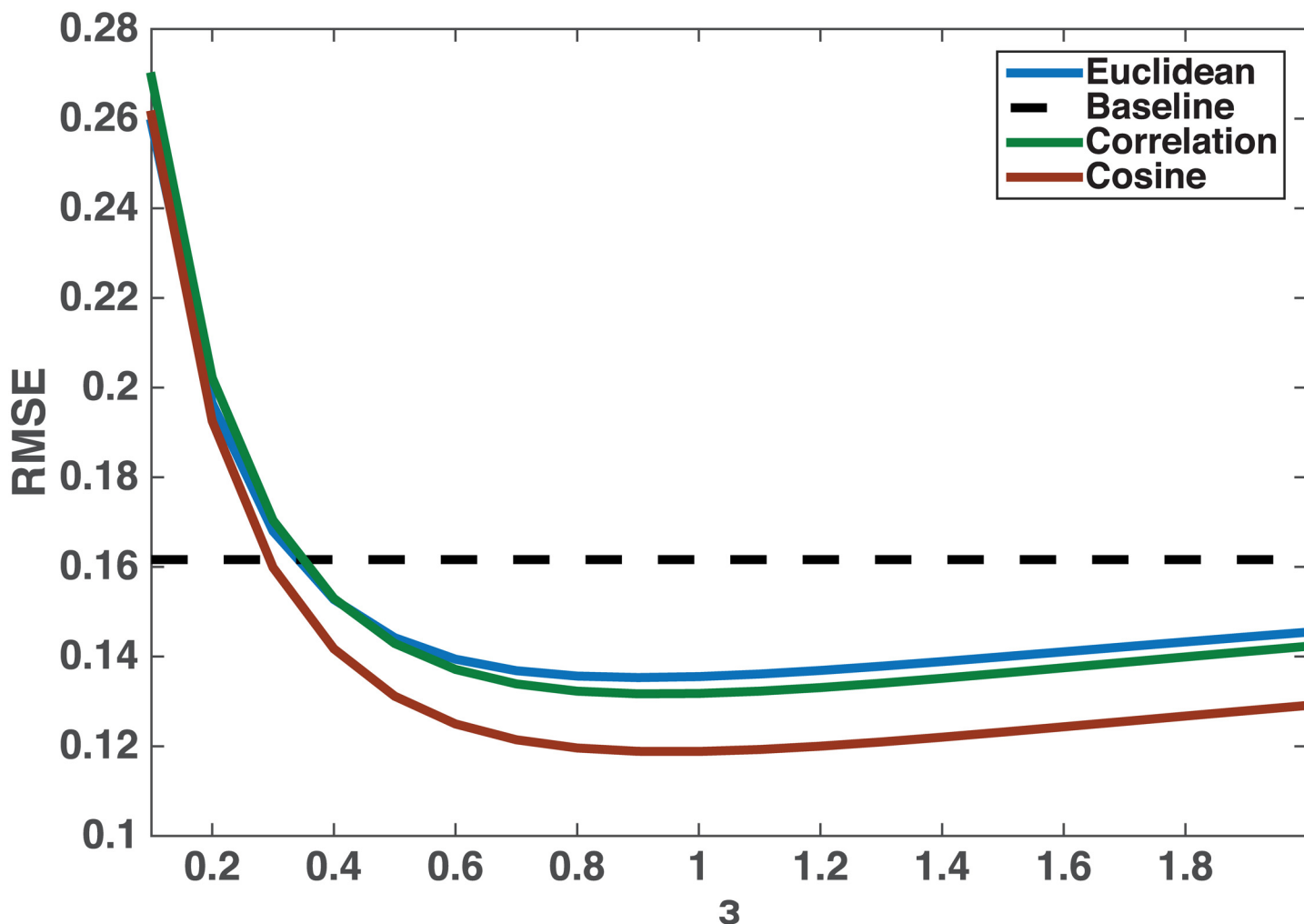


Fig 3. Sensitivity of our metric with respect to ξ and ϵ . The proposed VA-index outperforms the baseline extension of assortativity coefficient. Furthermore, it does not appear sensitive to the choice of ϵ (Eq (5)) and/or similarity metric.

doi:10.1371/journal.pone.0146188.g003

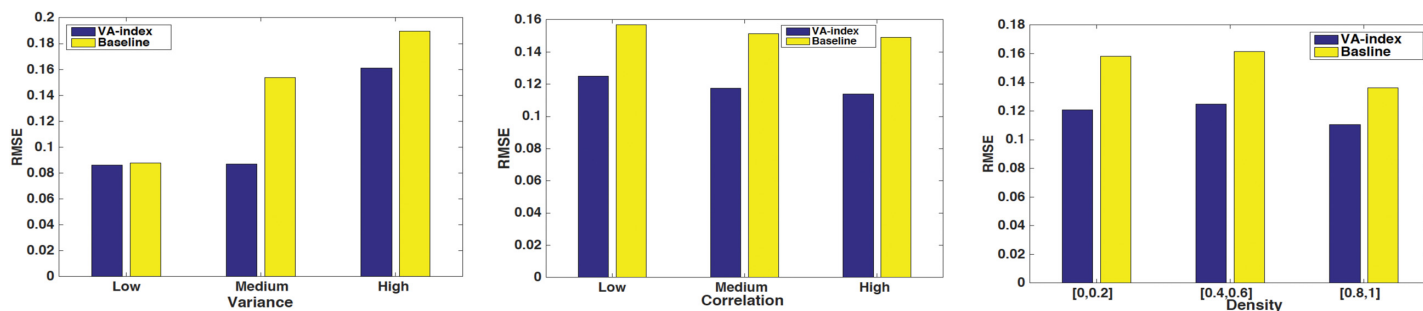


Fig 4. Comparison of the VA-index with the baseline extension of assortativity coefficient. The VA-index outperforms the baseline metric in all cases, irrespective of \mathbf{x} 's elements variance, correlation and the density δ of Σ . Nevertheless, for low variance the baseline performs almost equally as good with respect to the RMSE.

doi:10.1371/journal.pone.0146188.g004

Table 1. Mean difference Δe_v between the absolute error of our method and the baseline. The significance codes correspond to the two-sample t-test: 0 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 ' '. Low, medium and high density correspond to $\delta \in [0, 0.2]$, $\delta \in [0.4, 0.6]$ and $\delta \in [0.8, 1]$ respectively.

Dataset	ξ	Low	Medium	High
Variance	Cosine	0.0120***	-0.0691***	-0.0254***
	Euclidean	0.0067**	-0.0523***	-0.0228***
	Correlation	0.0064**	-0.0649***	-0.0254***
Correlation	Cosine	-0.0278***	-0.0267***	-0.0280***
	Euclidean	-0.0228***	-0.0230***	-0.0249***
	Correlation	-0.0267***	-0.0287***	-0.0316***
Density	Cosine	-0.02312***	-0.0282***	-0.0312***
	Euclidean	-0.0217***	-0.0262***	-0.0228***
	Correlation	-0.0267***	-0.0295***	-0.0311***

doi:10.1371/journal.pone.0146188.t001

two-sided t-test for the mean value of Δe , where the null hypothesis is $\mu_{\Delta e} = 0$, i.e., the two methods provide on average the same absolute error. If the null hypothesis is rejected, then the sign of $\mu_{\Delta e}$ will inform us which method provides smaller absolute error on average. Our results indicate that for all the three different similarity metrics we examined, $\mu_{\Delta e} < 0$, with a p -value < 0.01 . Table 1 further depicts our results split based on the variance and correlation of the vector elements and the density of Σ . More specifically, we present the mean value of Δe as well as the corresponding significance level. As we can see the VA-index always outperforms the baseline, except for the case of small variance where $\mu_{\Delta e} > 0$. However, in these cases the absolute value of $\mu_{\Delta e}$ is very small (close to 0) and one order of magnitude smaller compared to that for the rest of the cases where the VA-index outperforms the baseline. Furthermore, the significance levels of this difference are also smaller compared to the rest of the cases. Hence, we can conclude that our results imply that the VA-index is able to better recover the true assortativity of the network compared to a baseline extension of the assortativity coefficient.

Bias and Variance of the VA-index

Finally we examine the bias and the variance of the VA-index as an estimator.

Definition 0.1 Consider the real-valued statistic U for estimating a real number $\theta \in \mathbb{R}$. Then, we define as the bias of the estimator U , $bias(U)$, the difference between this estimator's expected value and the true value of the parameter being estimated, i.e.,

$$bias(U) = \mathbb{E}(U - \theta) = \mathbb{E}(U) - \theta \quad (9)$$

Based on the above definition, an unbiased estimator is one whose expected value is equal to the true value being estimated. An unbiased estimator is clearly a desired property. However, the variance of the estimator is another property whose value has implications on the quality of the estimation. With the mean square error of the estimator being $mse(U) = \mathbb{E}[(U - \theta)^2]$, we have for the variance, $var(U)$:

$$mse(U) = var(U) + bias^2(U) \quad (10)$$

Ideally we would like to have an unbiased estimator with small variance (i.e., small mean square error). However, this is not always possible and hence, we evaluate the performance of the VA-index with respect to the bias and variance as a function of its parameter ϵ . In particular, we generate 100 synthetic network topologies. We choose the Euclidean-based similarity metric to compute the corresponding VA-index, since this is the worst-case setting that gives the largest error (Fig 3). In order to be able to compute the empirical bias and variance of VA-

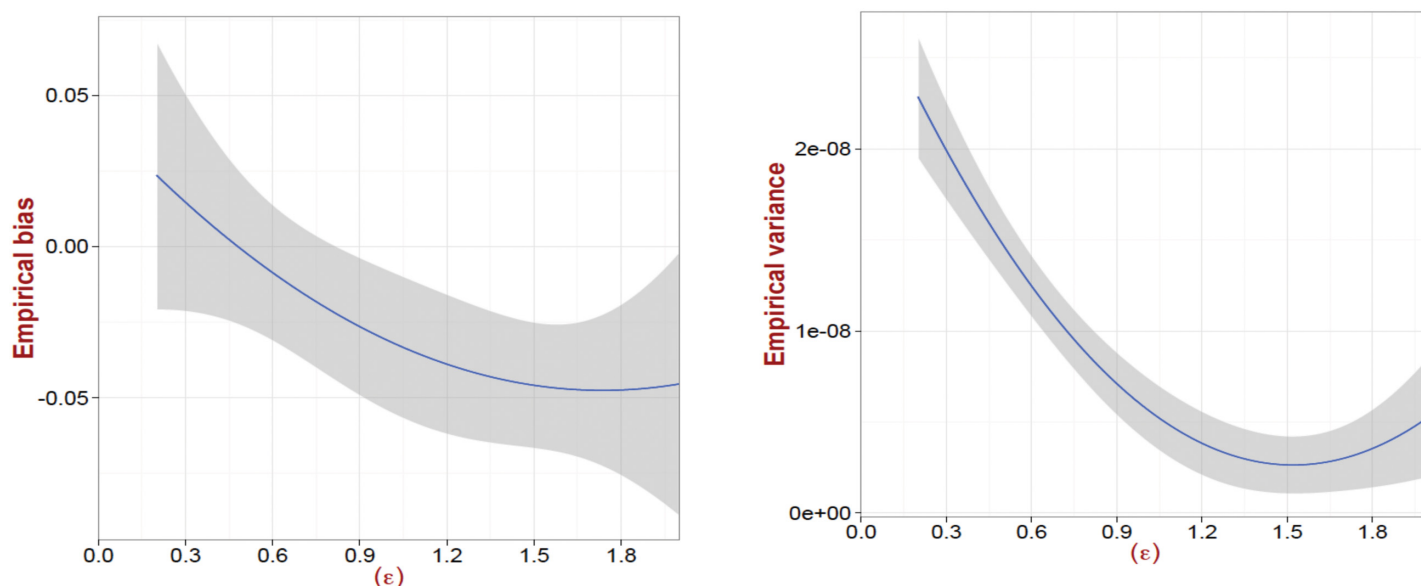


Fig 5. The bias and the variance of the VA-index. Both the bias and the variance of the VA-index have small absolute values. However, values around $\epsilon = 1$ appear to provide the best performance with regards to minimizing the mean square error of the estimator.

doi:10.1371/journal.pone.0146188.g005

index we perform this estimation 50 times (through 50 different applications of the bootstrap process) for each topology and value of ϵ . Fig 5 depicts our results. As we can see both the bias and variance of the VA-index are small (in absolute values). However, in the range (1, 2) for ϵ we see that the variance is “minimized”, while in the range (0, 1) the bias exhibits a small absolute value (as compared to the one in the range (1, 2)). Taking into consideration both the bias and the variance of the VA-index, values close to $\epsilon = 1$ appear to be appropriate for minimizing the mean square error all together, similar to what we identified above in Fig 3.

Application of VA-index on quantifying mobility assortativity patterns

Next we turn our attention to a real network dataset, and in particular, a dataset from a location-based social network (LBSN), namely, Gowalla, provided to us by Scellato *et al.* [23]. An LSBN consists of two components; (i) the social component that resembles any other digital social network, where users are connected based on “friendship” relations, and, (ii) the location component, which describes the mobility of the users based on their voluntary sharing of their whereabouts (through check-ins). Our dataset consists of 10,097,713 check-ins performed by 183,709 users in 1,470,727 distinct venues. Furthermore, there are 765,871 edges in the social (friendship) network.

Based on the above, every user u in this type of networks can be associated with a vector \mathbf{c}_u that captures the places he has visited. In particular, the i^{th} element of the vector is equal to the number of check-ins that u has in location/venue i . An important question that arises then is “What are the assortativity patterns of this network with respect to the mobility trails of the users?”. The answer to this question has implications for the underlying spatial homophily of this network [7, 24]. For answering this question we rely on the VA-index, where we use the cosine similarity as our similarity metric. In particular, the similarity between users u and v is defined as:

$$\xi_{u,v} = \frac{\mathbf{c}_u \cdot \mathbf{c}_v}{\|\mathbf{c}_u\|_2 \|\mathbf{c}_v\|_2} \quad (11)$$

Table 2. There is a clear positive assortativity mixing with regards to the mobility trails of Gowalla users. Even when controlling for the home-distance distribution the average pairwise similarity in the real network is significantly higher compared to that of a randomized network.

Real network similarity	ER network similarity	Controlled randomization
0.05425	[0.00233, 0.0024]	[0.01834, 0.01837]

doi:10.1371/journal.pone.0146188.t002

For our randomization we will consider two scenarios. First, we completely randomize the edges in the network, essentially sampling the $G(n,m)$ Erdős-Rényi random graph ensemble. Nevertheless, this will lead to an underestimation of the average pairwise similarity since the vast majority of (randomly selected) pairs will inevitably live in long distances and hence, the chances of having common venues visited will be small. Therefore, we will also perform a randomization where we will control for the distribution of the home-location distance of friends in the real network. Table 2 presents the computed average similarities for the real network as well as the 95% confidence interval from 100 instances of the two randomization processes. As we can notice the average pairwise similarity in the real network is significantly higher as compared to the one for the randomized networks. In particular, the average similarity in the real network is higher than the upper bound of the 95% confidence interval for both cases. It is also interesting to observe that the average similarity for the pure random graph network model is also significantly smaller as compared to the one in which we control for the home-location distance distribution of connected nodes.

We can then compute the $VA-index$, which is equal to 0.94 ($p\text{-value} < 0.05$), if we consider the pure ER network model as our baseline, and 0.31 ($p\text{-value} < 0.05$), if we control for the home-location distribution in our randomized baseline. As we can see the selection of the baseline (randomization) model is really important and is application specific. For example, in the scenario examined it is clear (for the reasons aforementioned) that the ER model overestimates the observed mixing patterns in the network.

Discussion

In this work we design an assortativity metric, $VA-index \alpha$, for multi-attributed networks. Our evaluations on synthetic data show that our metric can identify the mixing patterns of the network and outperforms a baseline extension of the assortativity coefficient. We believe that our work will not only trigger more research on this largely ignored to date topic but it will also drive the development of related metrics for composite networks. The latter can be thought of as multidimensional networks with multiple types of edges and nodes. In such networks a direct application of metrics developed for traditional (unimodal) networks will lead to a large information loss [26]. For example, as alluded to above, when there are multiple types of edges attached to a node, the degree of a node is not a scalar number but a vector that describes the number of different types of edges attached to the node. Hence, using the assortativity coefficient to calculate the degree mixing of this network will ignore significant amount of information. Nevertheless, the $VA-index$ will be able to take into consideration the various types of degree simultaneously and hence, provide a more accurate view of the degree assortativity in composite networks.

Supporting Information

S1 Text. Choice of the number of bootstrap samples \mathcal{B} .
(PDF)

S2 Text. Synthetic Network Generation.

(PDF)

S3 Text. Similarity metric ξ .

(PDF)

Author Contributions

Conceived and designed the experiments: KP. Performed the experiments: KP DW. Analyzed the data: KP DW. Contributed reagents/materials/analysis tools: KP. Wrote the paper: KP DW.

References

1. Newman MEJ. Mixing patterns in networks. *Phys. Rev. E* 2003; 67, 026126. doi: [10.1103/PhysRevE.67.026126](https://doi.org/10.1103/PhysRevE.67.026126)
2. McPherson M, Smith-Lovin L, Cook J. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 2001; 27:415–44. doi: [10.1146/annurev.soc.27.1.415](https://doi.org/10.1146/annurev.soc.27.1.415)
3. Bearman P, Moddy J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology* 2004; 110:44–91. doi: [10.1086/386272](https://doi.org/10.1086/386272)
4. Moddy J. Race, school integration, and friendship segregation in america. *American Journal of Sociology* 2001; 107:679–716. doi: [10.1086/338954](https://doi.org/10.1086/338954)
5. Catania J, Coates T, Kegelsa S, Fullilove M. The population-based amen (aids in multi-ethnic neighborhoods) study. *American Journal of Public Health* 1992; 82:284–287.
6. Newman MEJ. Mixing Patterns in Networks; 2002. Preprint. Available: arXiv:cond-mat/0209450v2. Accessed 22 December 2015.
7. Zhang K, Pelechris K. Understanding Spatial Homophily: The Case of Peer Influence and Social Selection. *ACM WWW* 2014.
8. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S. Feedback effects between similarity and social influence in online communities. *ACM SIGKDD* 2008.
9. Kim M, Leskovec J. Latent multi-group membership graph model. *ICML*; 2012.
10. Vuokko N, Terzi E. Reconstructing randomized social networks. *SIAM SDM*; 2010.
11. Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL. Human Mobility, Social Ties, and Link Prediction. *ACM SIGKDD* 2011.
12. Hunter D, Handcock M, Butts C, Goodreau S, Morris M. ERGM: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software* 2008; 23(3).
13. Newman MEJ. *Networks: An Introduction*. 1st ed. Oxford University Press; 2010.
14. Foster J, Foster D, Grassberger P, Paczuski M. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences* 2010; 107(24). doi: [10.1073/pnas.0912671107](https://doi.org/10.1073/pnas.0912671107)
15. Block P, Grund T. Multidimensional homophily in friendship networks. *Network Science* 2014; 2(2):189–212. doi: [10.1017/nws.2014.17](https://doi.org/10.1017/nws.2014.17) PMID: [25525503](https://pubmed.ncbi.nlm.nih.gov/25525503/)
16. Sánchez PI, Müller E, Laforet F, Keller F. Statistical Selection of Congruent Subspaces for Mining Attributed Graphs. *IEEE ICDM* 2013.
17. Pelechris K. Matching patterns in networks with multi-dimensional attributes: a machine learning approach. *Social Network Analysis and Mining* 2014; 4(1):1–11. doi: [10.1007/s13278-014-0188-2](https://doi.org/10.1007/s13278-014-0188-2)
18. Fisher L, Ness JWV. Admissible clustering procedures. *Biometrika* 1971; 58(1):91–104. doi: [10.1093/biomet/58.1.91](https://doi.org/10.1093/biomet/58.1.91)
19. Kleinberg J. An impossibility theorem for clustering. *NIPS* 2002.
20. Lauw H, Shafer J, Agrawal R, Ntoulas A. Homophily in the digital world: A LiveJournal case study. *IEEE Internet Computing* 2010; 14(2):15–23. doi: [10.1109/MIC.2010.25](https://doi.org/10.1109/MIC.2010.25)
21. Zhao K, Ngamassi L, Yen J, Maitland C, Tapia A. Assortativity patterns in multi-dimensional inter-organizational networks: a case study of the humanitarian relief sector. *SBP* 2010.
22. Erdős P, Rényi A. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 1960; 5:17–61.
23. Scellato S, Noulas A, Mascolo C. Exploiting Place Features in Link Prediction on Location-based Social Networks. *ACM SIGKDD* 2011.

24. Pelechrinis K, Krishnamurthy P. Location affiliation networks: bonding social and spatial information. ECML/PKDD 2012.
25. Bender EA, Canfield ER. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A* 1978; 24(3):296–307. doi: [10.1016/0097-3165\(78\)90059-6](https://doi.org/10.1016/0097-3165(78)90059-6)
26. Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations* 2013; 14(2):20–28. doi: [10.1145/2481244.2481248](https://doi.org/10.1145/2481244.2481248)